PNAS
Proceedings of the
National Academy of Sciences
of the United States of America

Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms

P. Jonathon Phillips, Amy N. Yates
National Institute of Standards and Technology

Ying Hu, Carina A. Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline Castro, Geraldine Jeckeln
The U of Texas at Dallas

Rajeev Ranjan, Swami Sankaranarayanan , Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa
U. of Maryland

David White
UNSW - Sydney

Alice J. O'Toole
The U of Texas at Dallas

PNAS Proceedings of the National Academy of Sciences of the United States of America

Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms

Thanks

- **All participants**
- **FISWG**
- **ENFSI-Face**
- **OSAC-Facial Identification**
- **International Summer School on Biometrics**

- **FBI**
- **IARPA**
- **NIJ**
- **Australian Research Council**

Open access and download for free at
http://www.pnas.org/content/115/24/6171

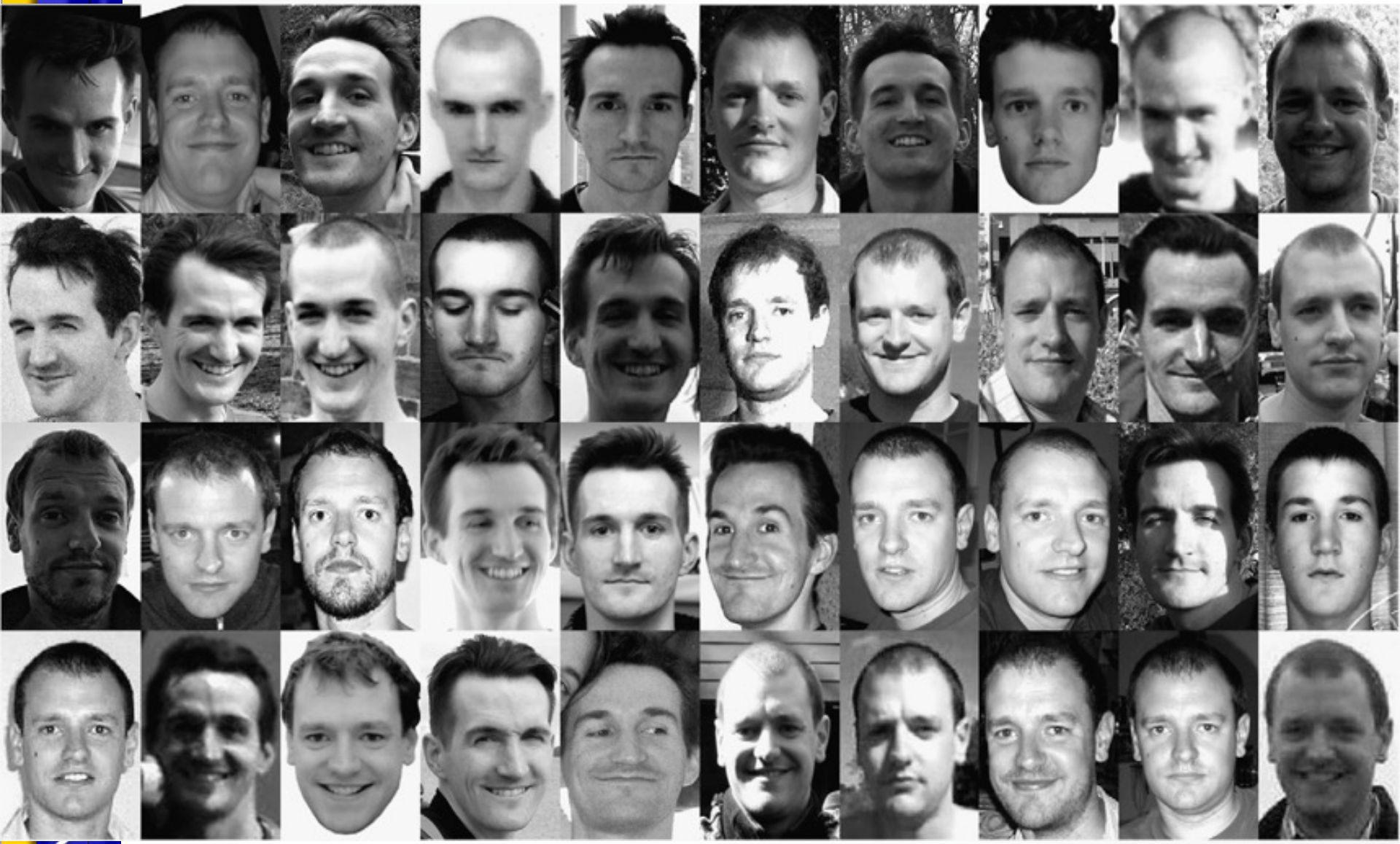
Scope of Study

- First to measure accuracy of facial examiners using tools and methods
- First to compare facial examiners and super-recognizers
- First to compare facial examiners and algorithms
- First to fuse facial examiners with algorithms

Pop Quiz

Who is this person?

How Many People?



Jenkins, White, Burton (2011)

What is a super-recognizer?

ELSEVIER
biometric TECHNOLOGY TODAY
ISSN 0969-4765 April 2016 www.biometrics-today.com

FEATURE

Humans vs machines: the future of facial recognition

Tim Ring, journalist

Tim Ring

London Metropolitan Police Commissioner Sir Bernard Hogan-Howe (left) presents the force's Staff of the Year award to 'super recogniser' Detention Officer Idris Bada.

Courtesy T. Ring, Biometric Technology Today, 2016

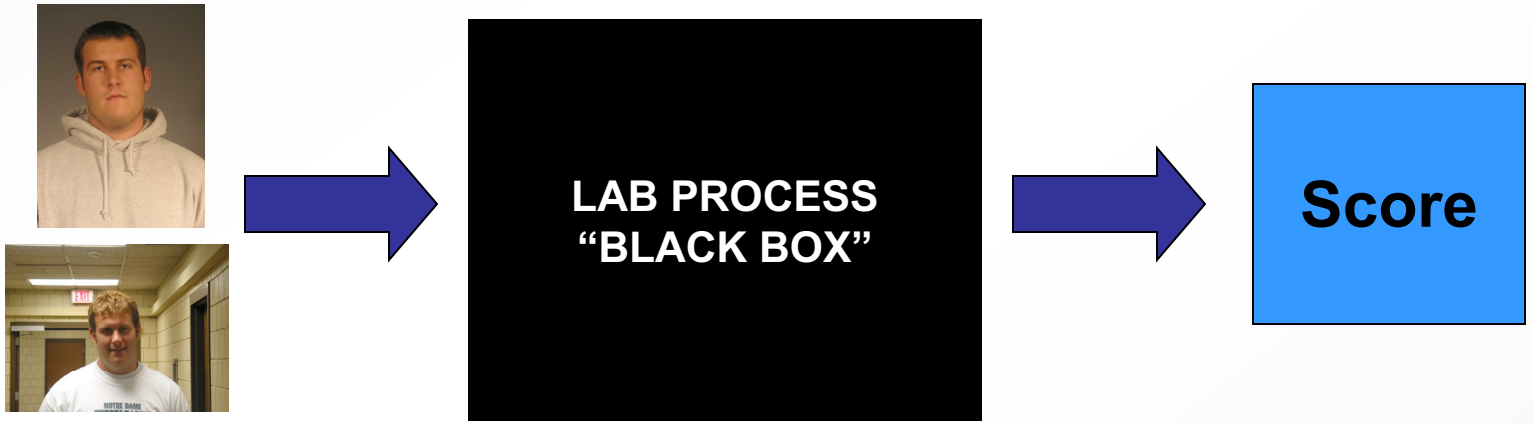
What is a facial examiner?

- Comprehensive comparison of faces in images
- Write detailed reports
- Prepared to testify in court
- Extensive training (2-4 years)

Credit: J. Stoughton/NIST

Black Box Study

- Measure performance of Forensic Facial Examiners *in situ*.



- Examiners were allowed access to lab procedures, tools, methods, resources, and time schedule (more or less).

General Rules

- 20 pairs of face images
  - Pre-screened by humans and machines to be ***extremely*** challenging
- 7 point comparison scale
- 3 months to complete comparisons

Same-identity pair

Different-identity pair

Selecting 20 Challenging Fair-pairs

9,307 images
570 subjects

Stratification By Algorithms

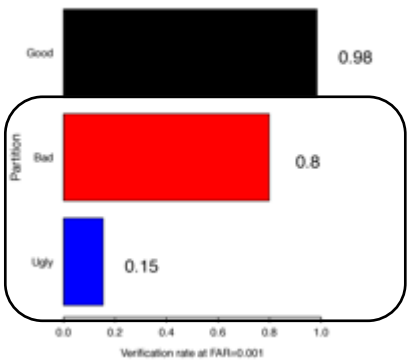
Select

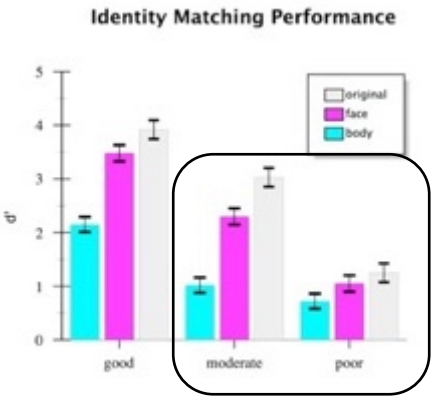
Student Human Performance

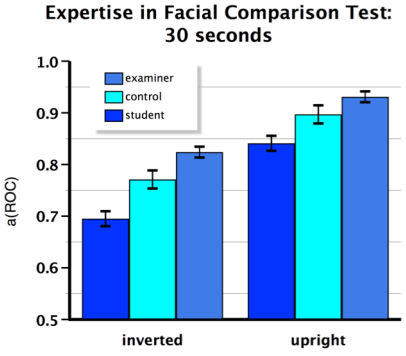
Select

Face Examiners 30 seconds

20 Extremely Challenging Face-pairs







Comparison / Identification / Matching

+3	The observations strongly support that it is the same person
+2	The observations support that it is the same person
+1	The observations support to some extent that it is the same person
0	The observations support neither that it is the same person nor that it is different persons
-1	The observations support to some extent that it is not the same person
-2	The observations support that it is not the same person
-3	The observations strongly support that it is not the same person

Four Subject Groups ++

- Facial forensic examiners (n=87, 5 continents)
  - Examiners (n=57)
  - Reviewers (n=30)
- Super-recognizers (n=13)
- Fingerprint examiners with no face experience (n=53)
- Undergraduate Students (n=30)

- **Algorithms**

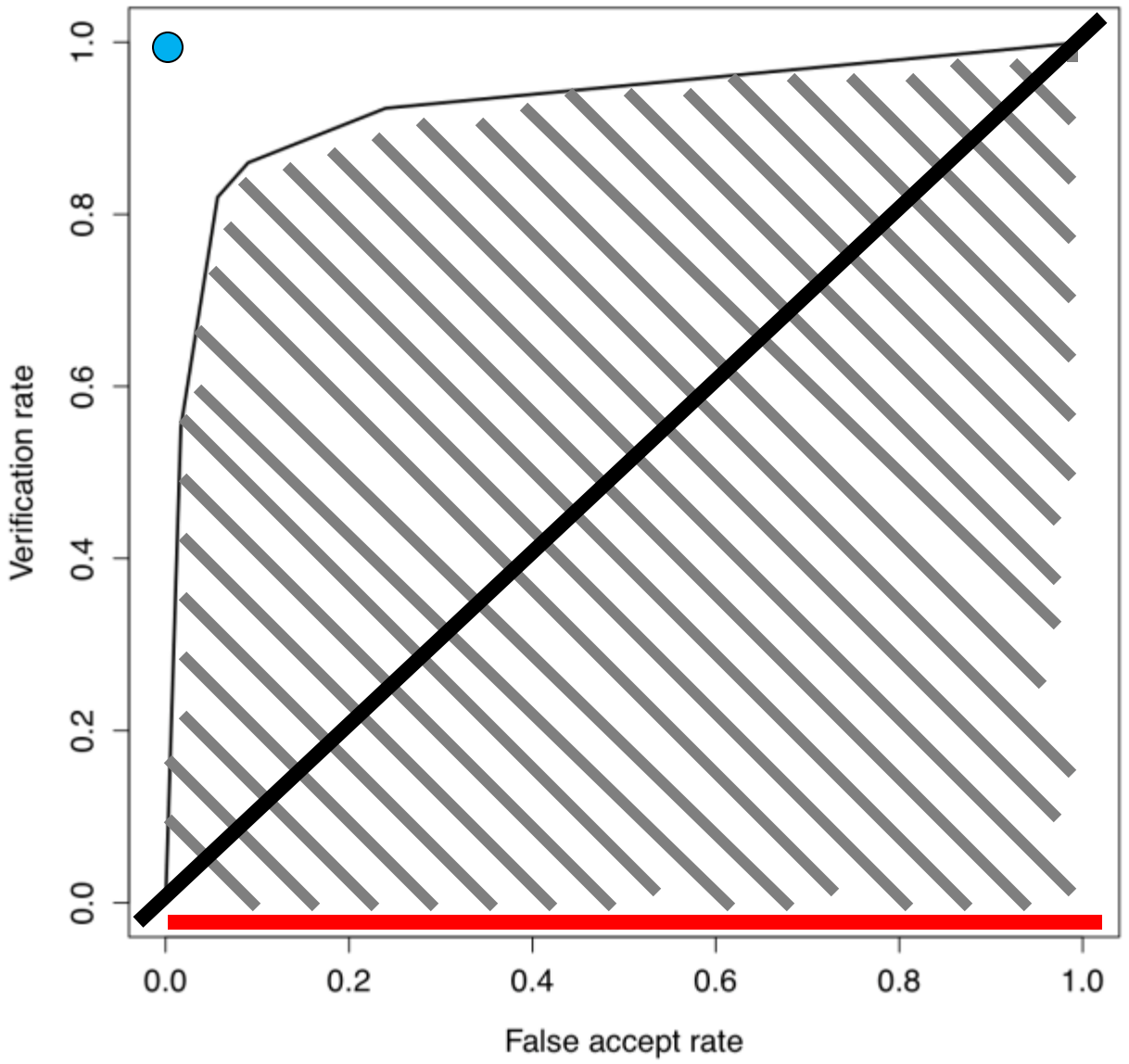
Algorithms

- VGG-Face (A2015)
  - Benchmark algorithm from Oxford
  - Deep convolutional neural network (DCNN) based
- U of Maryland
  - Rama Chellappa's group
  - 3 algorithms (A2016, A2017a, A2017b)

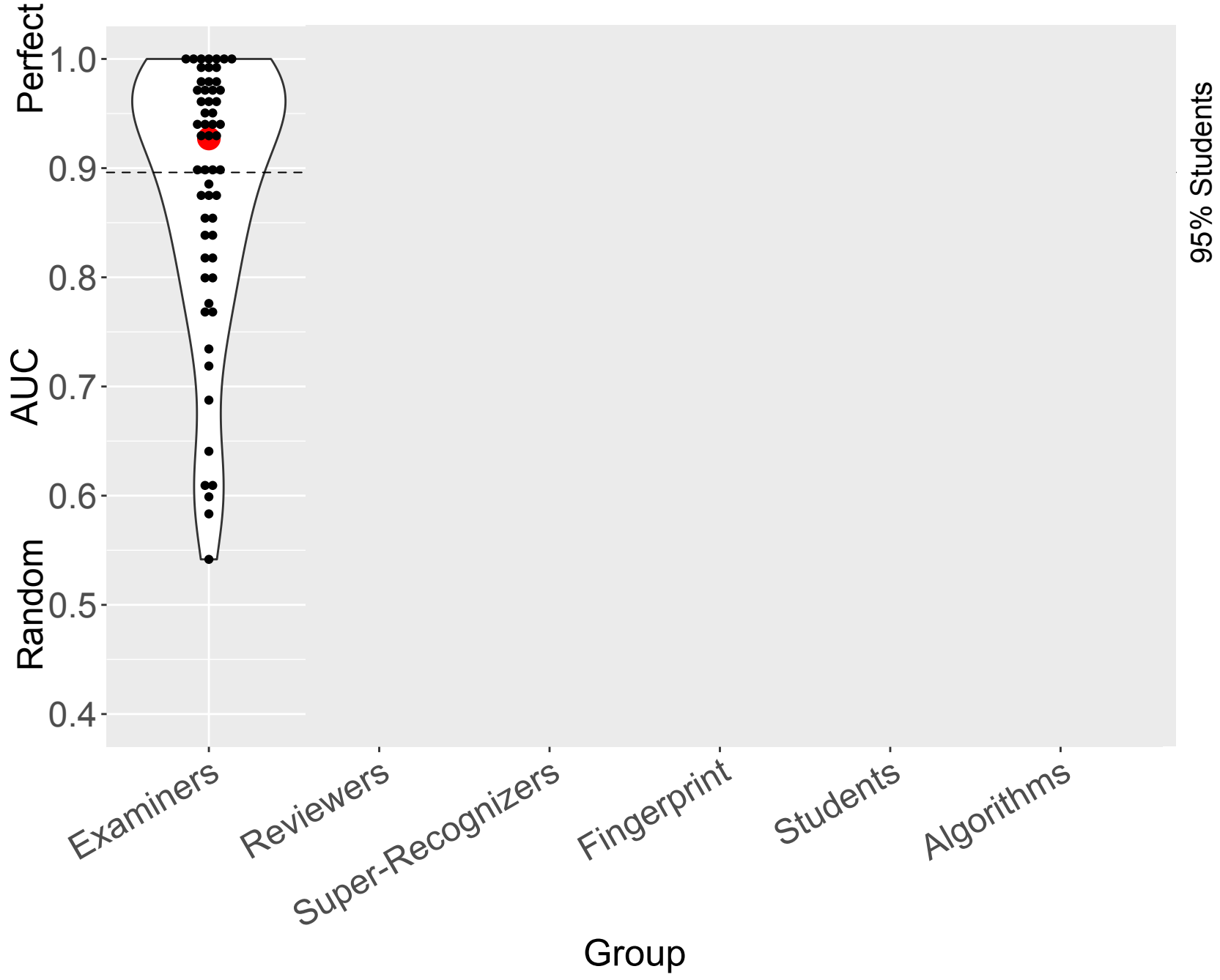
Four Major Questions

- Do facial examiners have superior ability?
- Is there a difference in accuracy between facial examiners and super-recognizers?
- How do algorithms compare to the humans with superior ability?
- Does fusion help?

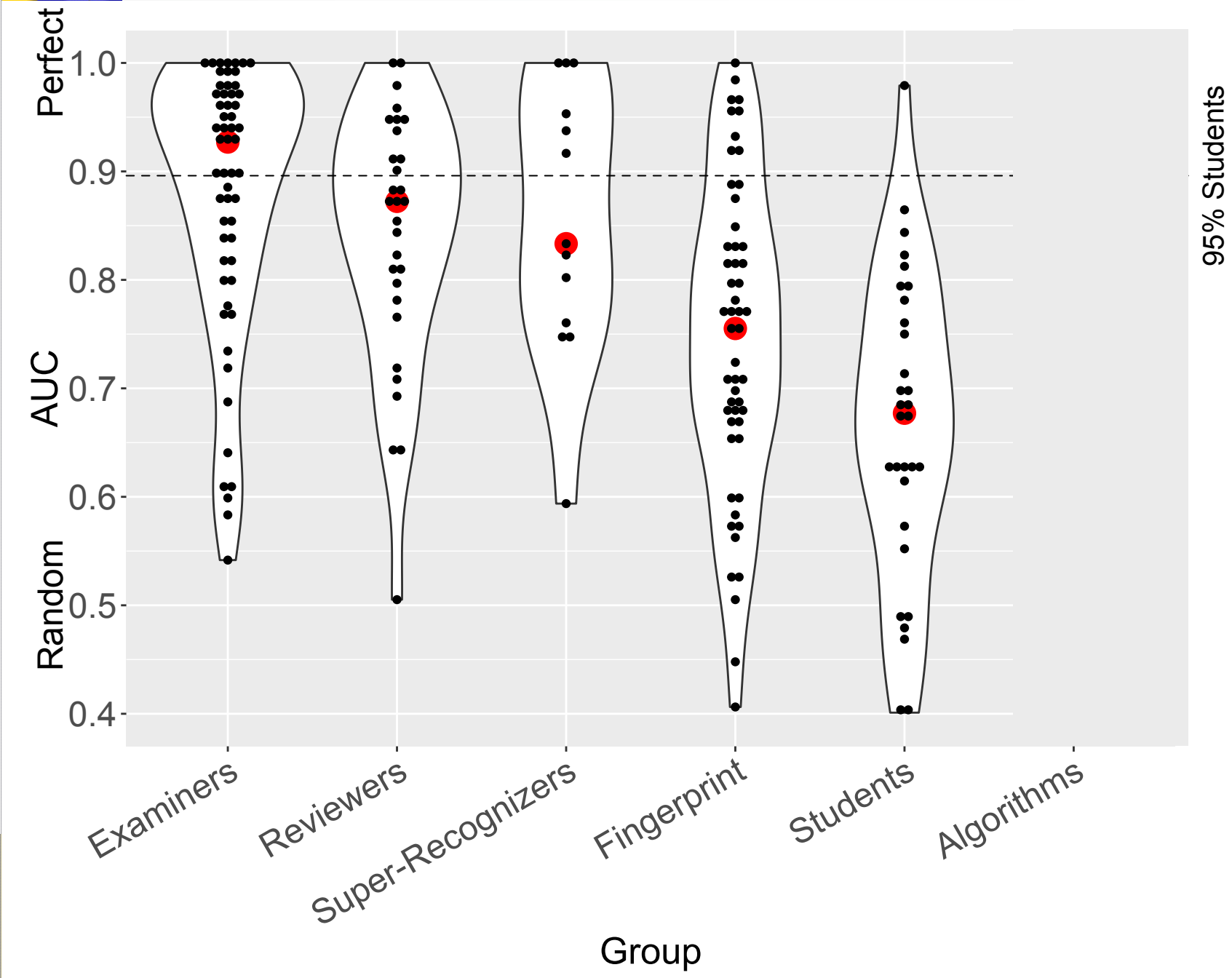
Area Under Curve (AUC)



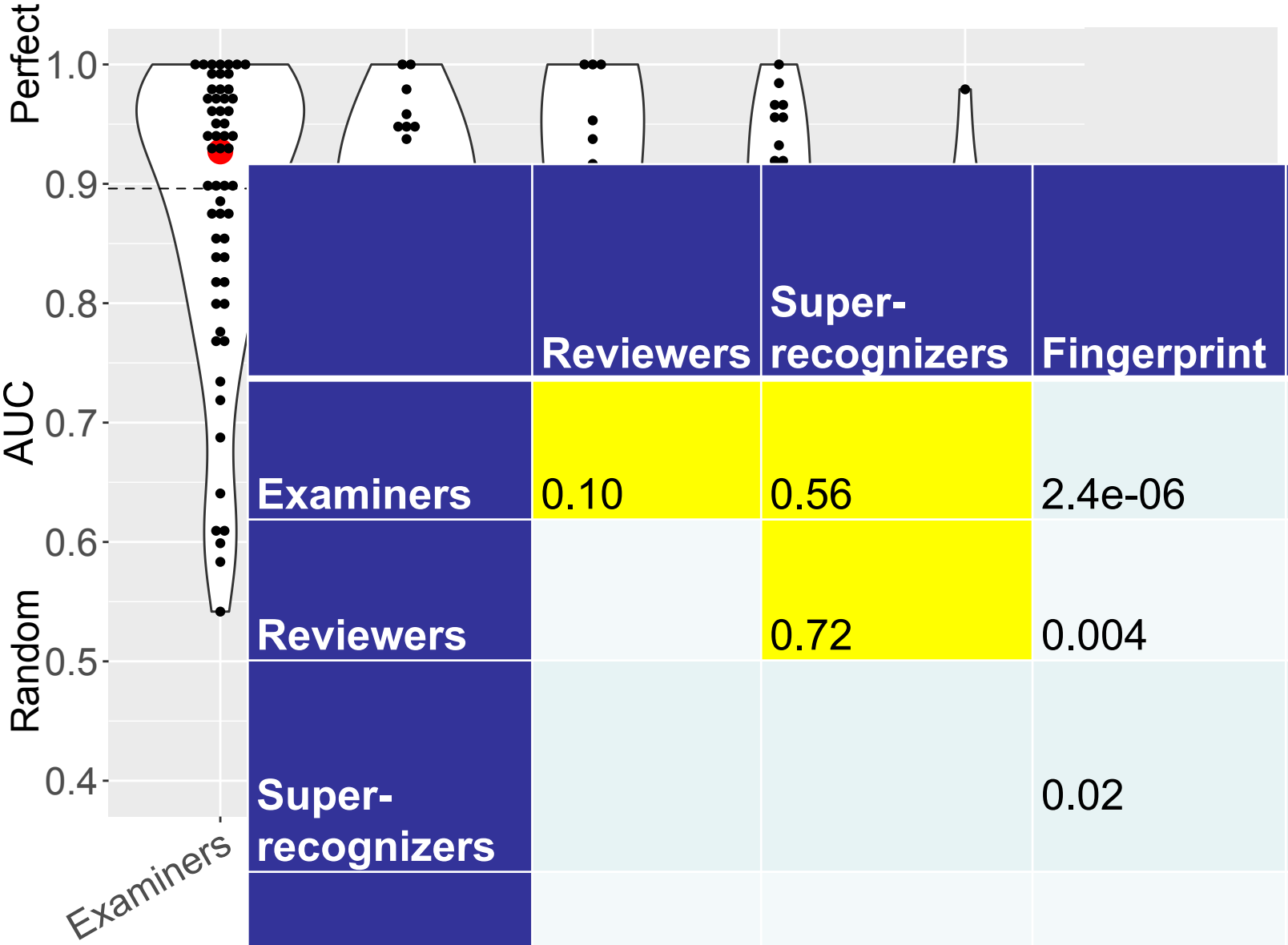
Comparison across subject groups and algorithms



Comparison across subject groups and algorithms

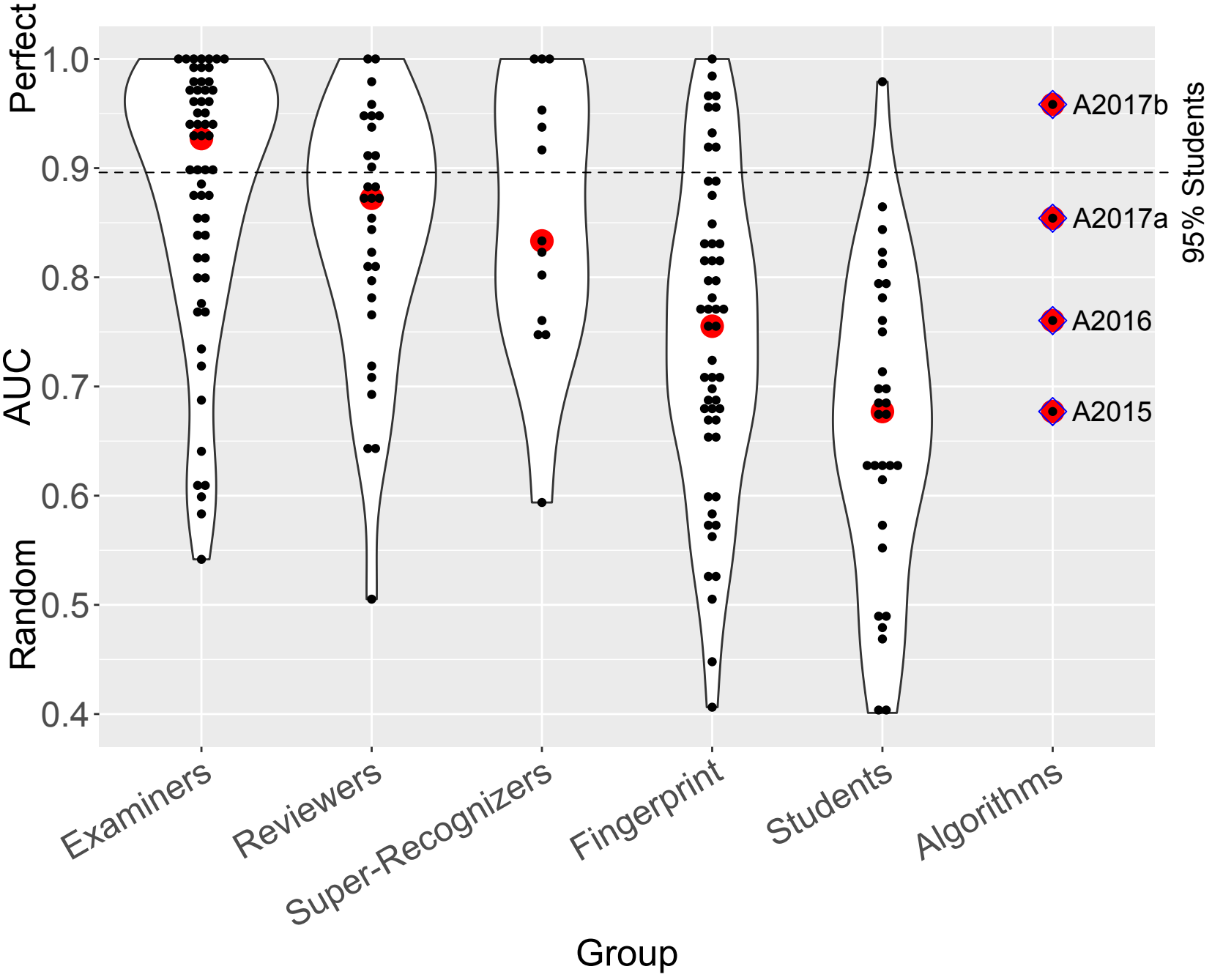


p-values for Mann-Whitney statistic

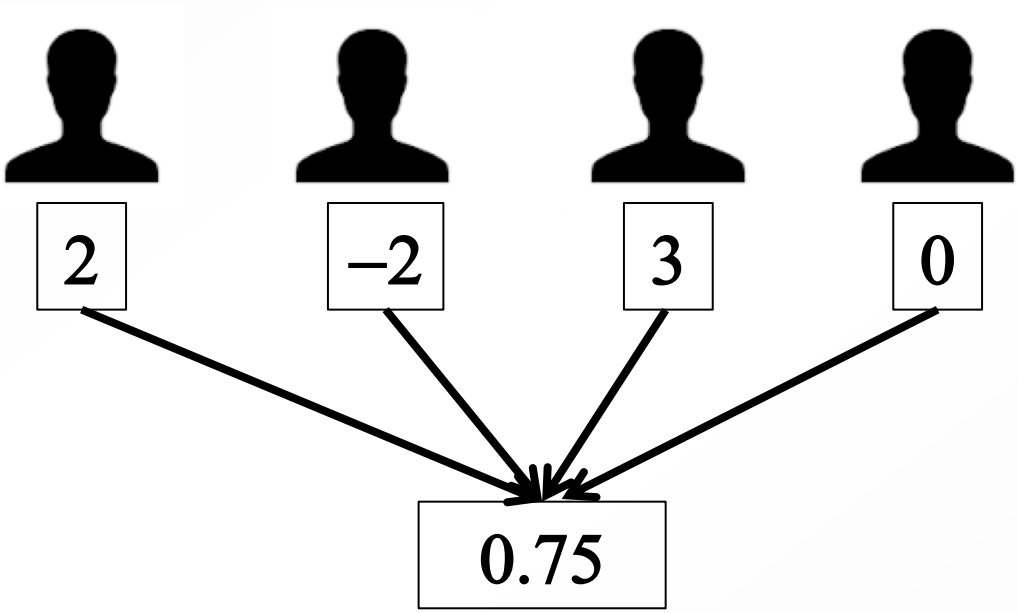


	Reviewers	Super-recognizers	Fingerprint	Students
Examiners	0.10	0.56	2.4e-06	2.5e-08
Reviewers		0.72	0.004	4.0e-06
Super-recognizers			0.02	0.0005
Fingerprint				0.02

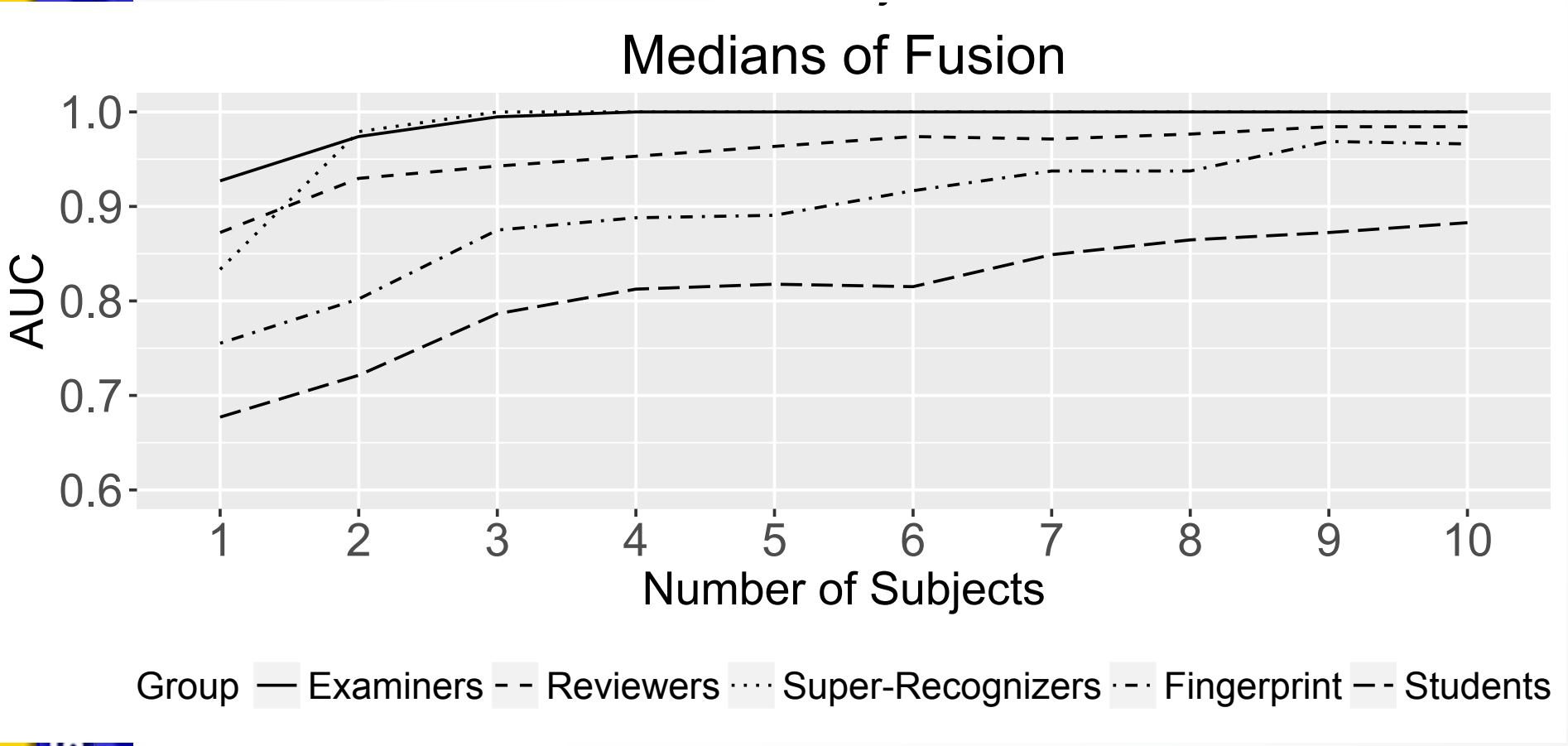
Comparison across subject groups and algorithms



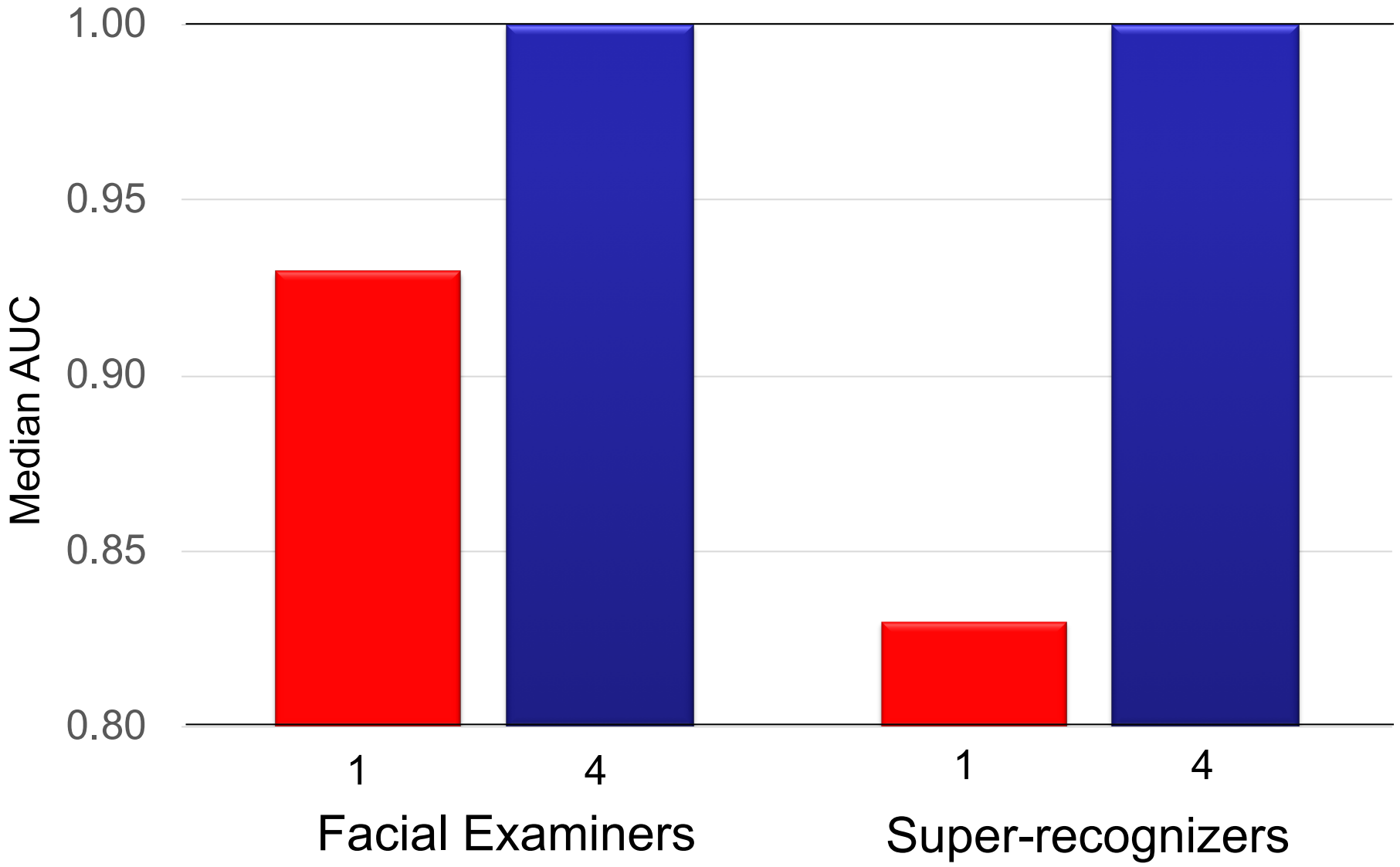
Independently Fusing Human Ratings



Effectiveness of Fusion

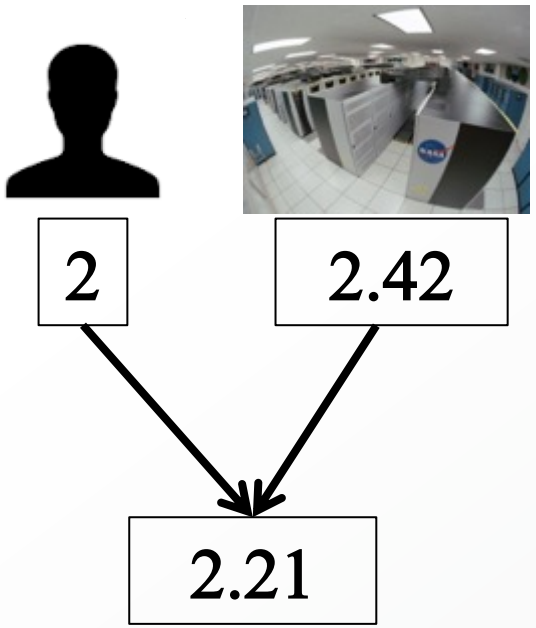


Fusing is Very Effective

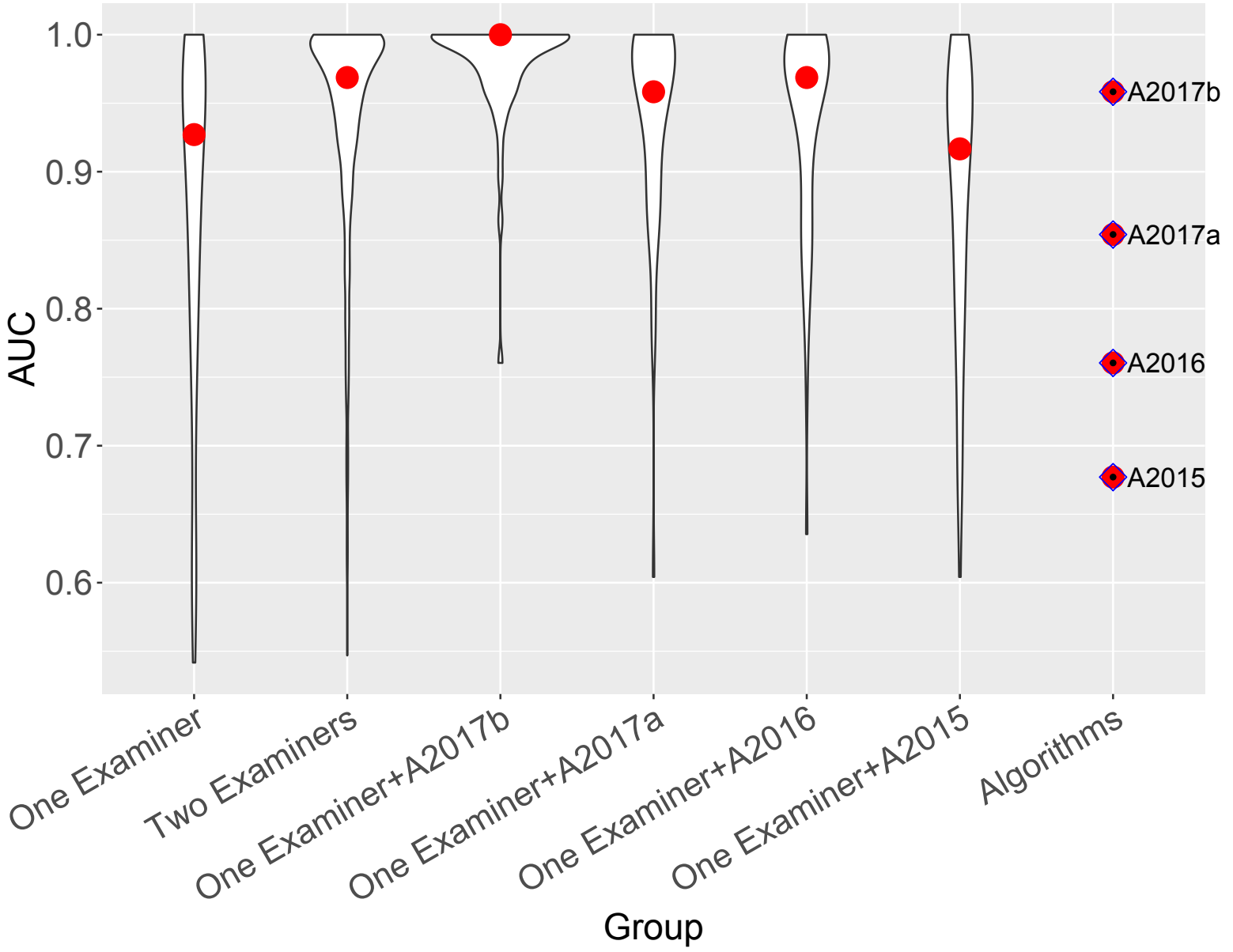


Can fusing examiners and algorithms improve accuracy?

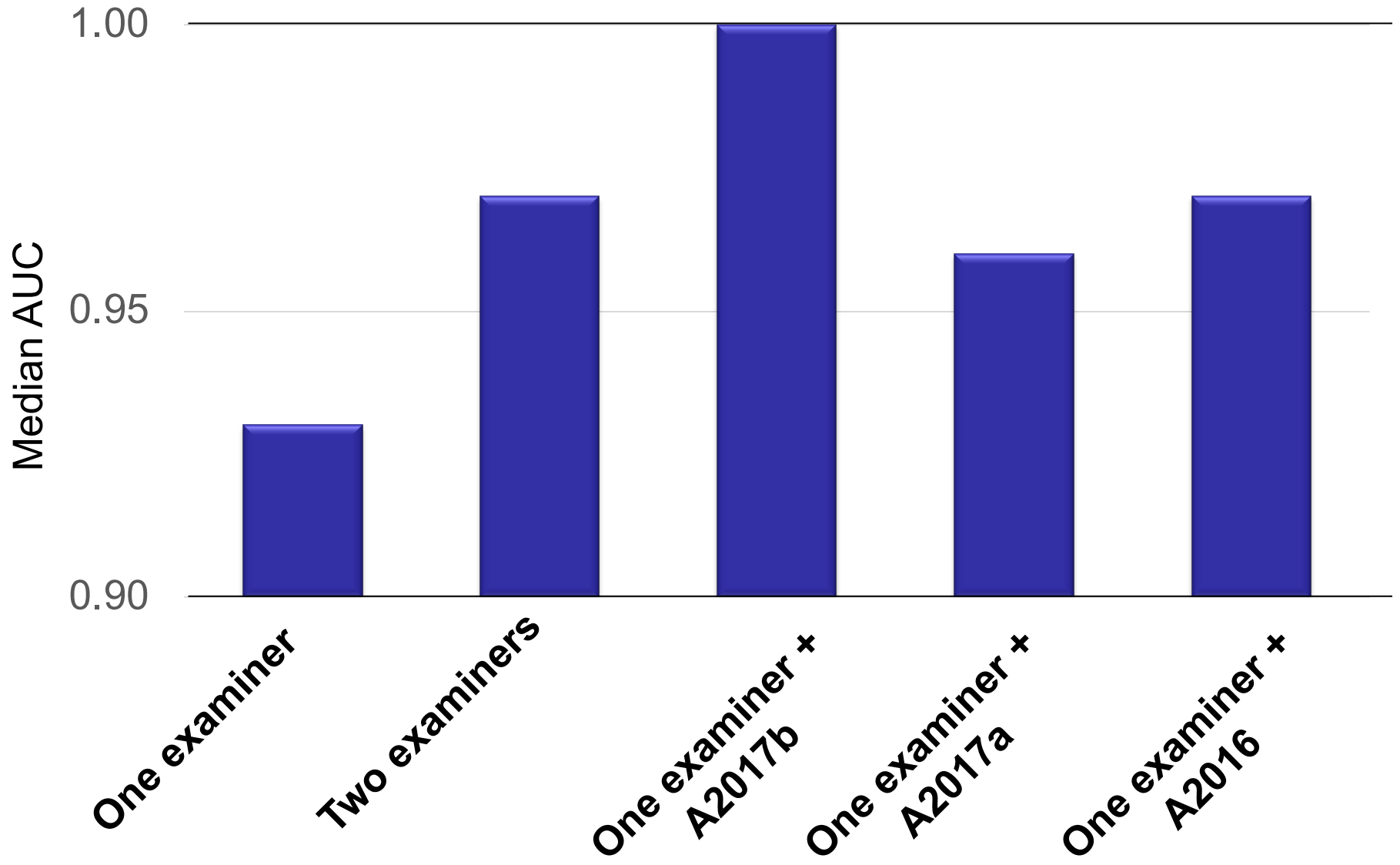
- Rescale algorithm scores
- Fusion by averaging



Fusing Examiners and Algorithms



Fusing Examiners and Algorithms



Conclusions

- Facial examiners are significantly better than the general population

- No statistical difference among examiners, reviewers, and super-recognizers.
- Best algorithm is competitive with best humans
- Fusing human judgements is effective
- Performance optimized by fusing one facial examiner and A2017b.

Next Steps

- Harder test of face recognition ability
- Accuracy across changes in
  - Pose
  - Blur
  - Video
  - Camera quality
- The other race effect

Future Goal

- Transitioning human-machine fusion to practice
  - Work with facial forensic community
  - Explaining algorithms decision
  - How it complements human decisions
  - Classic AI problem

General Audience Article

ELSEVIER
biometric TECHNOLOGY TODAY
ISSN 0969-4765 October 2018 www.biometricstoday.com

FEATURE

The great debate: study proves whether people or algorithms are best at facial ID


P Jonathon Phillips


Alice J O'Toole

P Jonathon Phillips, NIST and Alice J O'Toole, University of Texas at Dallas

Thank You